I Understand What You Are Saying: Leveraging Deep Learning Techniques for Aspect Based Sentiment Analysis

Jie Tao
Fairfield University
Fairfield, CT, USA, 06824
jtao@fairfield.edu

Lina Zhou
The University of North Carolina at Charlotte
Charlotte, NC, USA, 28223
lzhou8@uncc.edu

Conor Feeney
Fairfield University
Fairfield, CT, USA, 06824
conor.feeney@student.fairfield.edu

Abstract

Despite widespread use of online reviews in consumer purchase decision making, the potential value of online reviews in facilitating digital collaboration among product/service providers, consumers, and online retailers remains under explored. One of the significant barriers to realizing the above potential lies in the difficulty of understanding online reviews due to their sheer volume and free-text form. To promote digital collaborations, we investigate aspect based sentiment dynamics of online reviews by proposing a semi-supervised, deep learning facilitated analytical pipeline. This method leverages deep learning techniques for text representation and classification. Additionally, building on previous studies that address aspect extraction and sentiment identification in isolation, we address both aspects and sentiments analyses simultaneously. Further, this study presents a novel perspective to understanding the dynamics of aspect based sentiments by analyzing aspect based sentiment in time series. The findings of this study have significant implications with regards to digital collaborations among consumers, product/service providers and other stakeholders of online reviews.

1. Introduction

Online (consumer) reviews, as a type of electronic word-of-mouth, has become an important data source for various decision making processes. For instance, consumers may utilize online reviews of products or services to make purchasing or patronizing decisions. In addition, an online review platform can also serve as a platform for digital collaborations between consumers and businesses, between product/service providers and online retailers, and among different functional departments (e.g., marketing, sales, and manufacturing departments) within a business. For example, businesses may utilize online reviews as an alternative channel for marketing analysis to gauge consumers' perception and acceptance of their products. Online reviews also facilitate the communications between marketing and/or customer service department and manufacturers to identify areas of product/service improvements through extracting customer complaints and product defects from the contents of online reviews. Further, online reviews serve as an efficient channel for communications among peer customers. Online retailers may provide value-added services to businesses and consumers by offering helpfulness voting function and online recommendations [1]; and business operation and/or sales teams can employ online consumer reviews as a source of information for forecasting product sales [2].

Despite the potential of online reviews for digital collaboration, the sheer volume and free-text form of online reviews create significant barriers for insightful analysis. An increasing number of studies have focused on extracting sentiments toward target products from online reviews. However, extracting sentiments alone from online reviews is neither sufficient nor straightforward. First, customers do not always express their sentiments explicitly in online reviews. Second, even if sentiments can be extracted, their interpretations are context-dependent. For instance, "low (price)" and "low (quality)" are opposite in terms of the polarity of sentiments. As a result, traditional sentiment analysis methods are rendered "out-of-context" [3]. There is an emerging trend toward aspect based sentiment analysis (ABSA), which aims at identifying both aspects and their associated sentiments from review texts [4] , where *aspect* is used to refer to product/service attributes, functions, and parts. Nevertheless, aspect extraction from online reviews is a non-trivial task in itself. Third, studies have shown that using a sentiment lexicon (e.g. SentiWordNet [5]) may not effectively capture sentiments in online reviews. Therefore, to understand what reviewers are saying, it is important to examine sentiments toward

URI: https://hdl.handle.net/10125/59487
ISBN: 978-0-9981331-2-6

specific aspects of products/services, with the help of implicit semantic information from online review contents.

This study aims to address the limitations of previous studies in understanding online review content from the following aspects. First, previous research has examined sentiment analysis and aspect extraction in isolation [6] with few exceptions. Second, among the few studies that have investigated ABSA (e.g., [1] [7]), they first utilized either keyword lists (of aspects) or unsupervised probabilistic models (such as Latent Dirichlet Allocation, LDA) to extract aspects from documents, and then analyzed the sentiments towards those aspects (e.g., [5]). In other words, they take little account of possible interactions between aspects and sentiments. Third, they examined online reviews in static with little regard to their temporal dynamics.

In this study, we propose an analytical approach to ABSA that extends the state of research in three folds. First, we treat the extraction of aspects and sentiments as one holistic classification problem. Compared with the previous methods, our approach is able to not only automatically determine the labels for aspects and sentiment, but also identify the implicit relation(s) between aspects and sentiments embedded in texts. Second, our approach is able to capture the dynamics of aspect-based sentiments through analyzing time series of product/service features embedded in online reviews. To the best of our knowledge, this is the first study that examines aspect based sentiments in time series. Third, we propose a semi-supervised method for preparing the training datasets for building classification models, which would otherwise be a very time-consuming and labor-intensive process. Our proposed method leverages deep learning based natural language processing techniques. In addition, this study gains technical insights by empirically comparing different deep learning models and text preprocessing methods. Last but not least, the findings of this study provide managerial implications for businesses by applying the best-performed model to large datasets of online reviews we collected from Yelp.com.

The remainder of this paper is structured as follows. Section 2 summarizes related work on electronic word-of-mouth, Aspect Based Sentiment Analysis, and deep learning for text analytics. Section 3 introduces the study data and the analytical pipeline. Section 4 presents the analytical results; and Section 5 discusses the implications of our results from both the research and practice perspectives, and Section 6 concludes the paper.

2. Related Work

2.1. Electronic Word-of-Mouth

A large body of extant studies investigates how electronic Word-of-Mouth (eWoM) – also known as *online reviews* – affects future business performances, or the consumers' perceptions of them. Traditionally, researchers focus on the meta data (e.g., number of reviews, date of posting) of online reviews, and their relationship(s) toward business performances (e.g., future sales) [8]. However, the rich information embedded in the textual contents of the online reviews provides possibility of understanding the dynamics of purchasing/patronage decisions. In practice, customers may rely on online reviews to make their decisions. Thus, understanding the (textual) contents of online reviews provides value for understanding/predicting different characteristics of businesses (e.g., consumer perceptions). To this end, studies have examined the sentiment(s) expressed in online reviews, and their relationship to business performances. For instance, Chern et al. [6] classify online reviews by their polarity for the purpose of forecasting product sales. This study relies on a sentiment lexicon that is constructed by domain experts to derive weights of semantic categories in individual reviews; and then utilizes a Naïve Bayes classifier to classify sentences into different sentiment categories. Unlike other studies on sentiment analysis, the current research analyzes sentiments using sentiment signals (keywords) specific to different product types, in addition to generic sentiment signals.

2.2. Aspect Based Sentiment Analysis

Building on the studies as discussed above, a recent stream of research investigates sentiment signals specific to different aspects of products/services. In practice, customers pay varying levels of attentions to different aspects of products/services. For instance, a customer may select (or not select) a restaurant because of the food or the ambience. Siering et al. [1] study sentiments toward different features within online reviews of airline companies, for the purpose of explaining and predicting airline recommendations. However, this study relies on a manual approach to identify different features of airlines, which limits the coverage of identified features. Li et al. [7] investigate online reviews of tablet computers in terms of their features (processor, RAM, screen size). To identify different features, this study first utilize an unsupervised topic modeling technique; and then use supervised classification models to identify sentiment(s)

in online reviews. Similar method can be found in study [9]. Nevertheless, the aforementioned studies have not accounted for the dynamics of ABSA over time.

2.3. Deep Learning in Text Analytics

Traditional machine learning techniques (e.g., Naïve Bayes) have dominated text mining (e.g., sentiment analysis) for a long period. Yet textual data is often high dimensional and sparse – thus, researchers recently employ deep learning techniques for text analytics. Convolutional deep learning techniques is applied in ABSA [10]. Recurrent Neural Network (RNN) captures sequential and contextual information; RNN transits from one state to the next when the information is passed from one word to the next [11]. Given the inherent limitations of RNN (i.e. vanishing gradients and short dependencies), it is rarely applicable in real world problems. The latest development in RNN is long short-term memory (LSTM) models, which can capture long term dependencies and prevent exploding gradients with back-propagation [11].

Deep learning techniques, including LSTM, can be used in two phases in text analytics. First, they can be used in text representation for information retrieval and other similar tasks. For instance, Tsai et al. [12] utilize a continuous-space language model based on deep learning techniques to learn sentiment keywords in the finance domain. Also, deep learning techniques can be used to construct classification models in different scenarios. For instance, several deep learning based models are constructed, in comparison with traditional machine learning techniques, to predict stock movements based on the textual contents in financial reports [11].

3. An Analytical Pipeline for ABSA

To bridge the research gaps highlighted in Section 2, we design and follow an analytical pipeline, as shown in Figure 1.

The first step in the pipeline involves collecting the study data. We choose Yelp.com, one of the most popular online review platforms, as the source of data collection. In addition to reviews themselves, we also collect information about businesses. Given that the aspects of online reviews depend on specific products or services, we select restaurant as the industry for study.

The second step is data cleaning and merging. In addition, both traditional and deep learning based text preprocessing steps are conducted. Traditional text preprocessing tasks, such as sentence tokenization, stop word removal, stemming, lemmatization, and word filtering are (fully or partially) performed to prepare the datasets. More importantly, we follow a semi-supervised method to label the data, which are used to train and test the ABSA classification models. To assist the labeling process, we leverage an annotated dataset in the same domain from a different source.

In the ABSA classification phase, we train models using both traditional machine learning and deep learning techniques. In view of the multiclass classification nature of the ABSA problem, we select random selection as the baseline, and its classification accuracy is defined as the percentage distribution of the dominant class. The best performing models in terms of classification accuracy are selected and applied to new restaurant reviews.

Finally, we construct the feature time series by aggregating sentiments (positive, negative) and re-sampling the data on a monthly basis. These feature-sentiment time series are analyzed in reference to customer ratings, which is also operationalized as the monthly average star rating of the business. All analyses are conducted in Python. We introduce some of the key steps in detail next.

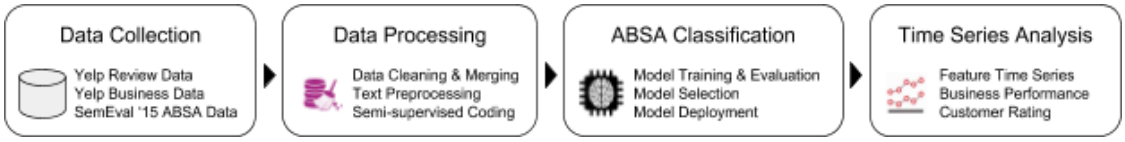


Figure 1. Analytical Pipeline for ABSA

3.1. Text Representation

Text representation is one of the fundamental task in processing text data (of online review texts). Word embedding is an emerging text representation method, where different linguistic units (e.g., words, sentences, paragraphs) are mapped to a space as high dimensional vectors [12]. We decide to select sentence as the unit of analysis in this study because it is a self-contained linguistic unit. Word embedding methods include discrete model based (i.e., label/one-hot encoding,

classic Bag-of-Words models) or continuous space model based (Continuous Bag-of-Words (CBOW), Skip-gram) [13], and pre-trained models (e.g., GloVe [14]).

Extant literature suggests that continuous-space model based word embedding better captures the contextual information [15][16]; and a recent study utilizes continuous space word embedding for learning domain related word lists [13]. Thus, we select continuous space word embedding as a preprocessing step in our semi-supervised data coding. Continuous word embedding is usually trained with deep learning models; in this study, we employed RNN models to train word embedding. Both CBOW and Skip-gram utilize the target word and its context, but their usages are different. CBOW predicts the target word given its context, whereas Skip-gram predicts the context given the target word. As a result, CBOW is often used to extend knowledge structures (e.g., finance word lists [12]), or to categorize textual data (e.g., clustering financial news [17]). As such, the Skip-gram model fits the purpose of this study better. A few key hyper-parameters of our skip-gram model are set as the following: the training going through 300 iterations (300 epochs), exerting a moderate control over "noise words" (negative sampling =5), and excluding words with total frequency less than 10. Table 1 presents the top 10 similar words of some sample aspects.

Table 1. Top 10 Similar Words from Our Skip-gram Model

Aspects	Ambience	Service	Quality
Top 10 Similar Words	noisy	food	food
	décor	great	price
	atmosphere	experiment	standard
	intimate	good	buffet
	pretty	quality	range
	park	price	nearly
	bright	ambience	great
	modern	attention	value
	elegant	restaurant	top
	deal	bartender	wynn

We can observe from Table 1 that: i) the results contain both words indicating specific aspects of restaurant and their associated sentiment signals; and ii) the results also capture words/sentiment signals of other aspects. Based on manual examination, we notice that a large number of sentences in the MAG reviews cover more than one aspect.

Given the obtained word embedding, we design a semi-supervised method to construct our training set. The method proceeds in layered fashion. In the first layer, we select 50 words that are most similar to our feature keywords. In the second layer, we select 30 most similar words for each of the selected words in the first layer based on the trained Skip-gram model. Then, we manually review all the candidate words, to construct the word list for each feature. The reviewed word lists are applied to a dataset consisting of all reviews on the top 10 restaurants in Las Vegas based on the number of reviews.

After aspect classification, we employ a pre-trained deep learning classifier to identify sentiments in each sentence from the review dataset. It is worth noting that, for each sentence in a review, we allow for overlap across different aspects, and sentiments over different aspects, but not across different sentiments over the same aspect. For instance, a sentence is removed if it contains both positive and negative sentiments towards the same feature *food*. In addition, we exclude a sentence if it does not express explicit sentiment toward a specific aspect.

3.2. Classification Models for ABSA

In this study, we train our classification models using both traditional machine learning and deep learning algorithms. The following configurations are used for traditional machine learning models. We use classic discrete BoW with term frequency – inverse document frequency (tf-idf) weighting in bi-gram representations, which is consistent with extant text classification studies [11]. Both linear and non-linear models are selected, namely Logistic Regression, Multinomial Naïve Bayes (MultiNB), and Support Vector Classifier (SVC). These models have been shown to be effective in text classification tasks [11], [18].

In terms of deep learning models, we select different network architectures, namely Multilayer Perceptron (MLP), Long Short-term Memory (LSTM), and bi-directional LSTM. MLP models are multi-layer logistic regressor in a feedforward network structure; yet it cannot handle sequential data. In contrast, both LSTM and bi-directional LSTM models are network structure with back-propagation as the optimization step; and they track sequential information by traversing from state s_n to s_{n+1}, when moving from word w_n to w_{n+1}. In addition, bi-directional LSTM track contextual information before and after any target word. Furthermore, LSTM advances RNN by capturing relatively long dependencies over input signals, by incorporating a cell state c_n - with both *remember* and *forget* gates. Thus, in addition to learning the weights over the word vectors in the obtained word embedding with the neurons, a vector of cell states is also learned during the training phase. After the last hidden layer, outputs are fed forward to an aggregation layer to make prediction/classification decisions. Since the nature of the LSTM model aligns closely with the goal of ABSA classification, we expect it to outperform other

alternative classifiers. Training both traditional machine learning and deep learning models requires tweaking their hyper-parameters. Previous related studies do not report on the details of model fine-tuning for deep learning classifiers, but only report certain combinations of hyper-parameters that are manually specified [12] [19]. In addition, we make the word embedding *train-able* in the modeling phase.

We select categorical cross entropy accuracy and log loss as the evaluation metrics for all models. Due to the multiclass classification nature of this study, we select categorical cross entropy accuracy to evaluation the classification accuracy, which is different from binary cross entropy accuracy used in binary classification models. In the context of multi-class classification, any model emits prediction $\hat{y} = h_\theta(x_i)$, in which x_i (i = 1, .. n) are predictors, and $h_\theta(x_i) = \sigma(Wx_i + b)$ is the function in each neuron. $\hat{y}$ is in form of a vector with a length of n (number of classes), and elements of probabilities (p_j) between 0 and 1. A certain class c is select as the predicted value of y iff. $\max(p_j|j = 1 \dots n) = p_c$. Thus, the baseline for comparison in multiclass classification is $1/n$, rather than 0.5 as in binary classification. In addition, for optimizing the model/classification performance, we use categorical cross entropy log loss, a commonly used optimization metric of the classification models – defined as follows:

$$-\log P(y_t|y_p) = -(y_t \log(y_p) + (1 - y_t) \log(1 - y_p)) \quad (1)$$

In which, y_t is the correct label of the target class, y_p is the prediction probability as emitted by the classifier. Log loss measures the inaccuracy of the prediction probability. By minimizing log loss, the classification model built using the labeled training sample can classify out-of-sample instances as accurate as possible.

All traditional machine learning models are implemented via the scikit-learn package. We use Keras as the frontend and TensorFlow as the backend to construct deep learning models.

We use a naïve baseline for model comparison, which randomly assigns a class to each sentence in the sample.

3.3. Aspect Based Sentiment over Time Series

Previous studies use content and meta data for ABSA [3] in the context of online reviews. Even though these methods can capture hidden semantic information embedded in online reviews with respect to aspects/sentiments, they overlook the dynamics of such aspects/sentiments pairs over time. In this study, we select the best performed model using approach and metrics discussed in Section 3.2, and then apply it on unlabeled sentences in online reviews to classify them into 12 classes. Then we resample the classified review sentences in terms of different time intervals (e.g., weeks, months, quarters), and quantify/aggregate them (e.g., use sentence counts, and aggregate counts from the same feature). The measurement of aspect *a* at time *t* is defined in equation (2):

$$a^t = \frac{a_p^t - a_n^t}{N_a^t} \quad (2)$$

where a_p^t and a_n^t denote the number of sentences concerning aspect a that express positive and negative sentiments, respectively, and N_a^t is the total number of sentences that discuss aspect a at time t. The time series of review aspects are analyzed in reference to review ratings.

4. Data Analyses and Results

In this section, we present the results of classification and time series analysis. We first report the comparative results of ABSA classification. Then, we demonstrate the usability of the identified aspects and sentiments via an exploratory time series analysis.

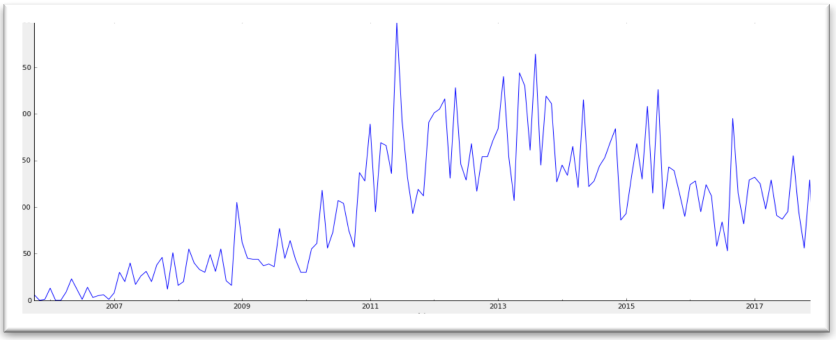
4.1. Data Analysis

The selection of aspects depends on the domain of online reviews. We selected restaurant as the target domain in this study for two main reasons. First, compared to online reviews of search goods (e.g., electronics), online reviews of experience goods such as restaurants have been less studied. Second, it remains difficult to find datasets that contain labels for both aspects and sentiments that are publicly available. Thus, the domain that has received more frequent reviews would be more preferred. The Yelp Dataset Challenge dataset consists of over 5 million online reviews and profile information of 174,000 businesses over 11 metropolitan areas that were under review. Based on our analysis of the business domains in the above online review dataset, restaurant is the most common domain.

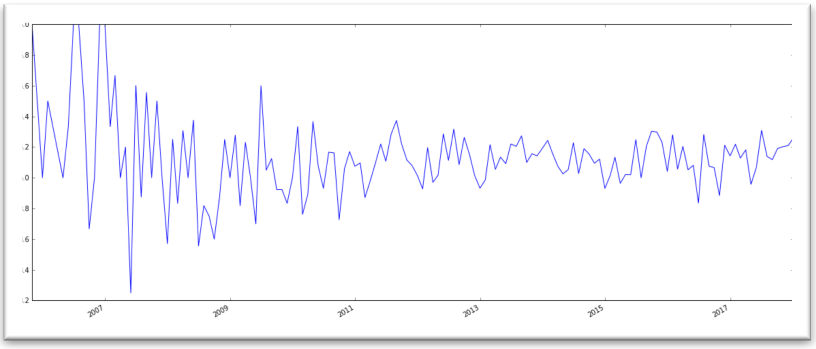
We determine the categories of restaurant aspects by adapting the data annotation schema used in SemEval '15, task 12 (http://alt.qcri.org/semeval2015/task12/). The original schema consists of 30 aspect-sentiment categories. In view that the dataset is composed of about 1,600 sentences, the large number of categories could lead to the over-specification problem for the subsequent classification task. To alleviate the problem, we group the aspects into 6 general categories: food (including drink), quality, price, service, ambience (including location), and

restaurant (i.e., miscellaneous). In addition, we exclude 'neutral' as a sentiment category in our study by assuming that analyzing neutral sentences would have little material effect on understanding business performance. Accordingly, the size of our label set used consists of 12 categories (aspect-sentiment pairs).

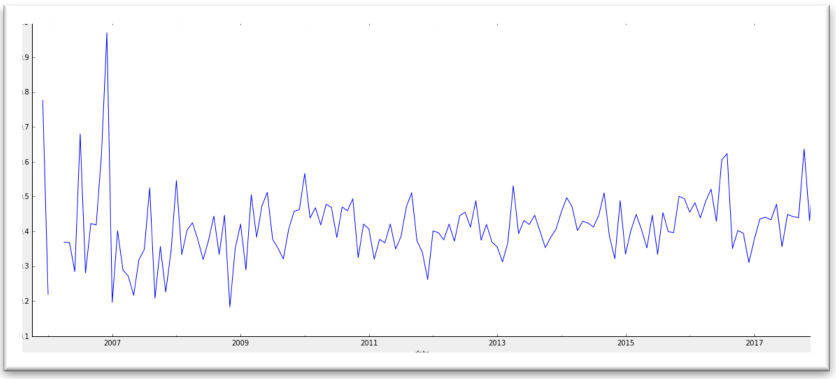
We randomly sample 2,000 sentences for each of the 12 feature-sentiment classes, and manually review them. These sentences are manually reviewed, which result in 10,951 sentences that are validated. Those sentences, along with 1,601 sentences from the SemEval '15 Task 12 dataset, are merged to support further data analyses.



(a) Review Count



(b) Star Rating



(c) Sentence-level Sentiment

Figure 2. Time Series of Monthly Average of MAG Reviews

We can observe from Figure 2 that the monthly average number of reviews of a selected restaurant (MAG) ranges from 0 – 300. And the monthly average star rating fluctuated prior to 2009 (possibly due to the limited number of reviews) – and converged around 4.3 stars in a 5-star rating system. Further, we examined the sentence-level sentiment of MAG reviews, which is operationalized as the average polarity score. It is shown in Figure 2(c) that the sentence-level sentiment of MAG reviews ranges from 0.3 to 0.6 after 2009, which indicates that the reviews toward MAG are generally *positive*. To further understand the content of MAG reviews, we also conduct content analysis using word clouds. Figure 3 (a-b) shows word clouds from MAG reviews at different time points, which confirms that the reviews of MAG are indeed generally *positive*.

We use a 90%/10% training/test split – our training dataset contains 11,296 sentences, and our testing dataset contains 1,256 sentences. We decide to use the categorical cross entropy accuracy and log loss of the validation dataset to select the best performing model, thus, we further conduct a 90%/10% split to get a validation dataset within the training set. Finally, our training dataset contains 10,166 sentences, our validation dataset contains 1,130 sentences, and our test dataset contains 1,256 sentences. We employ the same random state to ensure that the same split is used across different model configurations.

(a). May 2011

(b). June 2015

Figure 3. Word Clouds of MAG Reviews

4.2. Classification Results

The baseline accuracy is 18.82%, when all the sentences are assigned to the dominant class (FOOD#POSITIVE,). To support model comparison, we report for each model performance improvements in classification accuracy over the baseline.

Panel (a) in Table 2 summarizes the classification results using traditional machine learning algorithms – in which, Support Vector Classifier (SVC) achieves the best training accuracy/log loss, whereas logistic regression achieves the best validation accuracy/log loss. Panel (b) in Table 2 presents the ABSA classification results from using different deep learning network architectures and preprocessing methods. Among all deep learning based models, the fine-tuned LSTM model within the *complete preprocess + Skip-gram* configuration achieves the best test accuracy at about 50% (with an absolute improvement of 32.9% over the naïve baseline) and a log loss of 0.179. Thus, this model is selected to classify the unlabeled sentences for constructing the aspect time series.

Table 2. Comparative Results of ABSA Classifiers

Model Configurations Algorithm	Test Accuracy	Absolute Improvement over the Baseline
(a) Traditional Machine Learning		
LR	0.4713	0.2866
SVC	0.3877	0.2030
MNB	0.4307	0.2460
(b) Deep Learning		
(i) No Stemming/Lemmatization + CBOW		
MLP	0.3917	0.2070
LSTM	0.4323	0.2476
Bi-directional LSTM	0.4020	0.2173
(ii) No Stemming/Lemmatization + Skip-gram		
MLP	0.3464	0.1617
LSTM	0.4307	0.2460
Bi-directional LSTM	0.3936	0.2089
(iii) Complete Preprocessing + CBOW		
MLP	0.3981	0.2134
LSTM	0.4549	0.2702
Bi-directional LSTM	0.4371	0.2524
(iv) Complete Preprocessing + Skip-gram		
MLP	0.3806	0.1959
LSTM	**0.4998**	**0.3151**
Bi-directional LSTM	0.4363	0.2516

4.3. An Exploratory Time Series Analysis

With the best model selected, we use it to classify out-of-sample unlabeled sentences from all reviews of our selected restaurant (MAG) in our dataset (Yelp reviews). Figure 4 depicts the results of exploratory time series analysis. It is shown from the figure that

'RESTAURANT' accounts for most reviews among all 6 aspects, followed by 'QUALITY'. In contrast, 'PRICE' and 'SERVICE' account for lower ratios of reviews than other aspects. As far as aspect-sentiment pairs (e.g., "FOOD#POSITIVE")) are concerned, aspects 'RESTAURANT', 'QUALITY', 'FOOD', and 'LOCATION' are more inclined to the *positive* sentiment (above zero); while 'PRICE' and 'SERVICE' fluctuate between *positive* and *negative* sentiments – despite that its general sentiment is *positive*.

In order to make star rating time series comparable to aspect time series, we normalize the values of start rating into the range of [0,1]. , We normalize a star level sl_t at time t as: $sl_t = {(s_t - 3)}/{5 - 3}$, where s is the raw average star rating at t, and 3 is the mid-point in a 5 star system.

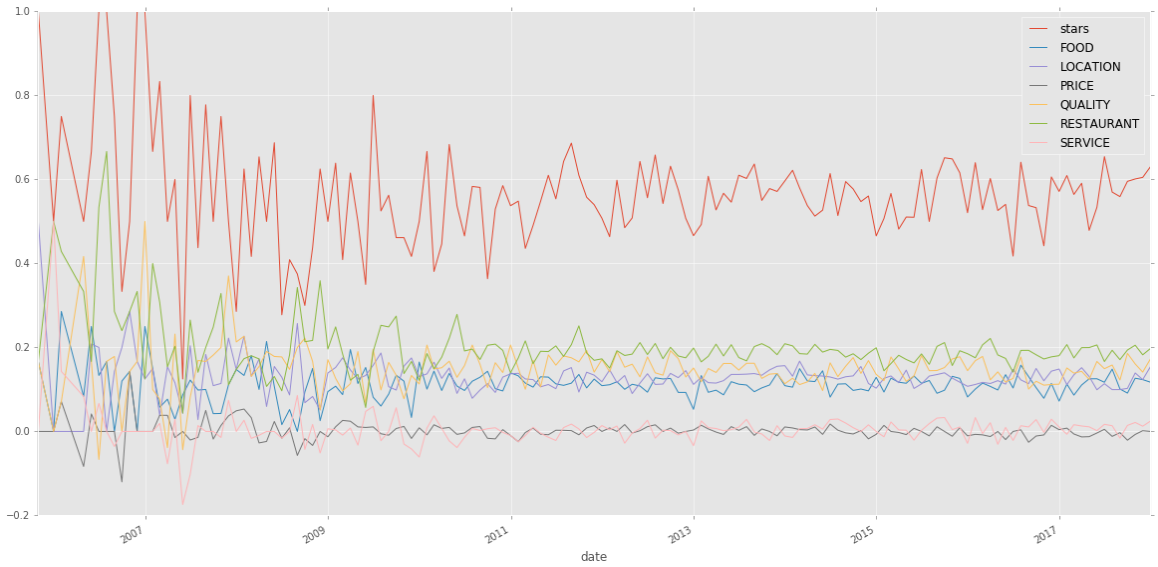


Figure 4. Time Series of Restaurant Aspects and Average Star Rating by Month

In order to quantify the similarity between different time series, we employ widely used similarity measures, including *dynamic time warping* (DTWDistance) and *Keogh lower bound* (LB Keogh) [20]. Both metrics measure how different a pair of time series are – the lower the values, the more similar the two series are. The similarity of each of the six aspect time series to the star time series are presented in Table 3.

Table 3. Similarity of Restaurant Aspect to Star Rating Time Series

Metric	FOOD	SERVICE	QUALITY	PRICE	RESTAURANT	LOCATION
DTWDistance	5.6322	6.7806	5.6649	6.7981	5.6135	5.8848
LB Keogh	5.4318	6.5724	5.4784	6.6849	5.3760	5.7218

Table 3 shows that 'FOOD', 'QUALITY', and 'RESTAURANT' are strongly correlated with the monthly average star ratings of MAG; whereas 'LOCATION' have relatively strong impacts. Surprisingly, 'PRICE' and 'SERVICE' features have relatively weak impacts on the monthly average star level. Based on manual inspection of the classification results, a relatively large number of sentences is misclassified between 'QUALITY#POSITIVE' and 'FOOD#POSITIVE', which partly explains why 'QUALITY' is the most important aspect rather than 'FOOD'. In addition, 'LOCATION' is the aspect with the highest classification accuracy (70.83%). A separate investigation reveals that the MAG restaurant is located across the street from a famous landmark in Las Vegas. In addition, customers tend to discuss less about 'PRICE' and 'SERVICE' of MAG, which partly explains why those aspects are weakly correlated with the monthly average star level.

Similar trends can be observed for the time series of aspects and business performance (ranking in terms of average star rating). It is also worth noting that the lowest point in the 'SERVICE' time series is in line with that in the 'star' time series – which indicates that service is an important factor in negative reviews for MAG.

5. Discussion

The findings of this study point to new directions for ABSA research, as well as suggest ways to improve the methods as introduced in our analytical pipeline. First, compared to traditional two-step (i.e., topic modeling and sentiment classification) approach, our proposed ABSA method utilizes the power of supervised learning – which can be more efficient and accurate. On the other hand, the classification performance can be improved by employing a larger, more balanced training sample; and by minimizing overlapping sentences among different aspect classes. Second, methodologically, we prove the value of text preprocessing, particularly stemming and lemmatization, in (multi-class) text classification. In addition, in line with previous studies [12], [18], the Skip-gram word embedding appears to be better suited for text representation for classification. Thirdly, as far as continuous-space word embedding is concerned, we find that excluding rare words (i.e., words with lower frequency) can improve classification results. Further, customizing word embedding in building deep learning classification models would contribute to improved classification performance.

The results from this study also provide practical insights for businesses by helping them respond to aspects that concern customers. If the average sentiment of certain aspect is negative within certain time period, and the average star rating also declines during the same time interval; the business should plan to improve that particular aspect to a satisfactory level. In addition, methods and results introduced in this study also enable the prediction of consumer perception (average star rating) of businesses in future using aspect time series as predictors. As an immediate next step, we plan to further improve the model for ABSA classification and construct aspect time series on a large sample of restaurants; and then we can use these time series to forecast/predict not only star ratings, but business performances and survival probability as well. We also plan to explore fusing the information extracted from the multimedia data of online reviews to better understand what consumers are saying.

6. Concluding Remarks

The effective use of online consumers' reviews to facilitate digital collaborations between consumers and product/service providers is contingent upon how well we could understand the review contents. Extracting aspects or sentiments from online reviews alone does not provide a complete picture of consumers' experiences and preferences with products/services. We investigate aspect-based sentiment analysis by conducting an experiment with online restaurant reviews. Specifically, we identify 12 aspect based sentiment categories based on content analysis of unsupervised machine learning results, propose a semi-supervised method for labelling online review contents with aspect based sentiments, and showcase how time series analysis can not only reveal the temporal dynamics of aspect-based sentiments but also shed light on the determining factors in consumers' rating of products and services. The results show that the time series of *food*, *quality*, and *restaurant* have a relatively higher similarity than *location* to those of review ratings, and the latter further has a higher similarity than *price* and *service* to those of review ratings from our experiment results.

Our experiment results suggest that deep learning techniques outperform traditional machine learning techniques in classifying online review contents with aspect based sentiments. In addition, the skip-gram model for learning text representations from review text led to better performance than its CBOW counterpart, and preprocessing textual contents with stemming and lemmatization can help boost the performance of extracting aspect based sentiments from them.

With increasingly widespread use of online consumer reviews, the analysis of aspect based sentiments paves the way for building an ecological system for businesses to improve customer relationship management and gain competitive advantages.

7. References

[1] M. Siering, A. V. Deokar, and C. Janze, "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews," *Decision Support Systems*, vol. 107, pp. 52–63, 2018.

[2] T. L. Ngo-Ye and A. P. Sinha, "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model," *Decision Support Systems*, vol. 61, pp. 47–58, May 2014.

[3] R. Y. K. Lau, C. Li, and S. S. Y. Liao, "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis," *Decision Support Systems*, vol. 65, pp. 80–94, May 2014.

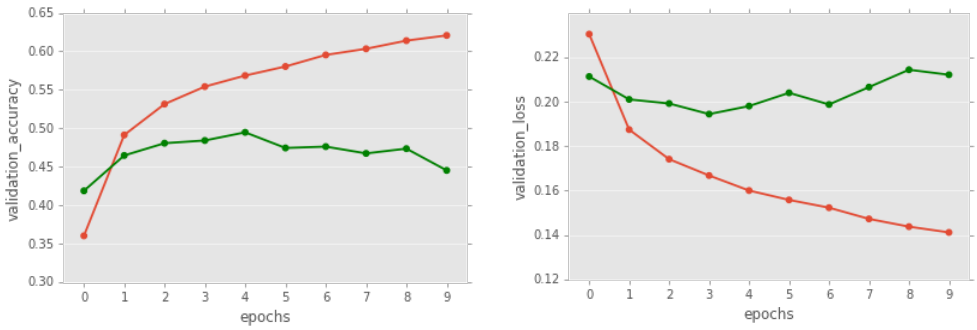
[4] Z. Hai, G. Cong, K. Chang, P. Cheng, and C. Miao, "Analyzing sentiments in one go: A supervised joint topic modeling approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1172–1185, 2017.

[5] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet," in *LREC*, 2010, vol. 10, pp. 2200–2204.

[6] C.-C. Chern, C.-P. Wei, F.-Y. Shen, and Y.-N. Fan, "A sales forecasting model for consumer products based on the

influence of online word-of-mouth," *Information Systems and e-Business Management*, vol. 13, no. 3, pp. 445–473, 2015.
[7] X. Li, C. Wu, and F. Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," *Information and Management*, no. April, 2018.
[8] N. Hu, N. S. Koh, and S. K. Reddy, "Ratings Lead You To The Product , Reviews Help You Clinch It : The Dynamics and Impact of Online Review Sentiments on Products Sales The Mediating Role of Online Review Sentiments on Product Sales," *Decision Support Systems*, vol. 57, pp. 42–53, 2013.
[9] R. Y. K. Lau, D. Song, Y. Li, T. C. H. Cheung, and J. Hao, "Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 6, pp. 800–813, 2009.
[10] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.
[11] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decision Support Systems*, vol. 104, pp. 38–48, 2017.
[12] M.-F. Tsai, C.-J. Wang, and P.-C. Chien, "Discovering Finance Keywords via Continuous-Space Language Models," *ACM Transactions on Management Information Systems*, vol. 7, no. 3, pp. 1–17, 2016.
[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.
[14] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," in *EMNLP. Association for Computational Linguistics*, 2013, pp. 1532–1543.
[15] Y. Li, B. Wei, Y. Liu, L. Yao, H. Chen, J. Yu, and W. Zhu, "Incorporating Knowledge into neural network for text representation," *Expert Systems with Applications*, vol. 96, pp. 103–114, 2017.
[16] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings," *ACM Transactions on Information Systems*, vol. 36, no. 2, pp. 1–30, 2017.
[17] Z. Hu, W. Liu, J. Bian, X. Liu, and T.-Y. Liu, "Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction," 2017.
[18] W. Zhao, Z. Guan, L. Chen, X. He, D. Cai, B. Wang, and Q. Wang, "Weakly-supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 185–197, 2018.
[19] B. Shi, G. Poghosyan, G. Ifrim, and N. Hurley, "Hashtagger+: Efficient High-Coverage Social Tagging of Streaming News," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 43–58, 2017.
[20] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping," *Transactions on Knowledge Discovery from Data (TKDD*, vol. 7, no. 3, pp. 3047–3051, 2013.

Appendix A



Note: Fine-tuned parameters and hyper-parameters of the selected model: i) network architecture: Input (dims. = 200) – Embedding –LSTM (128 neurons) – Dropout (rate = 0.3) – LSTM (128 neurons) – Dropout (rate = 0.3) – Dense (32 neurons) – Dense (Dims. = 12, activation = 'softmax'; ii) (hyper-)parameters: loss function: categorical cross entropy, optimization function: Adam(lr=0.001, clipnorm=.25, beta_1=0.7, beta_2=0.99); evaluation metric: categorical cross entropy accuracy.